Chapitre 2 : régression linéaire multiple

1 Le modèle

On considère maintenant que l'on veut prédire une variable Y à l'aide de plusieurs covariables numériques $X_1, \ldots, X_p$. Un **individu** de la population est donc représenté par le vecteur $(X_1, \ldots, X_p, Y)$. On suppose que

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

où $\varepsilon \perp\!\!\!\perp (X_1, \ldots, X_p)$, $\mathbb{E}(\varepsilon) = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$. Les paramètres de ce modèle sont $\beta_0, \ldots, \beta_p, \sigma^2$.

Comme dans le modèle de régression linéaire simple, on a

Proposition 1.

$$\mathbb{E}\Big(Y \Big| X_{1:p} = x_{1:p}\Big) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

En revanche, l'**interprétation** des paramètres est plus délicate (mais toujours faisable) :

— σ^2 : variabilité entre individus lorsque les covariables sont fixées,

— β_0 : prédiction ponctuelle de Y lorsque $x_1 = \cdots = x_p = 0$,

— $\beta_j, j \geq 1$: quantité à ajouter à la prédiction lorsqu'on augmente x_j d'une unité, **sachant que toutes les autres covariables restent fixées**.

Il est remarquable (et c'est une hypothèse du modèle) que cette dernière quantité ne dépend pas des valeurs des autres covariables. **Attention**, lorsque $p > 1$, le signe de β_j n'a pas de lien direct avec le signe de la corrélation entre X_j et Y. En fait, on a :

Proposition 2. *Pour* $j = 1, \ldots, p$,

$$\beta_j = \frac{\mathrm{Cov}\left(X_j, Y \middle| X_{(-j)} = x_{(-j)}\right)}{\mathrm{Var}\left(X_j \middle| X_{(-j)} = x_{(-j)}\right)}, \quad \text{où } X_{(-j)} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p).$$

Imaginons un exemple où Y est la masse graisseuse, et X_1 et X_2 représente la taille et le poids d'un être humain. Alors, X_1 et X_2 ont une corrélation positive

importante, et on peut avoir :

$$\mathrm{Cov}(X_1, Y) > 0,\ \mathrm{Cov}(X_2, Y) > 0 \quad \text{et} \quad \beta_1 < 0,\ \beta_2 > 0.$$

L'interprétation des effets β_j est donc toujours liée aux autres covariables présentes dans le modèle. On écrit parfois l'expression « **toute chose étant égale par ailleurs** », l'augmentation de X_j d'une unité induit une augmentation de Y de β_j en moyenne. Il faut faire attention au fait que l'expression entre guillemets cache les covariables que l'on a mis dans le modèle avec X_j. Si on change les covariables (que l'on enlève ou en ajoute une), la valeur de β_j, de même que son interprétation va changer.

On a décomposé Y en somme de deux variables indépendantes $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ et ε. On a donc

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\beta_1 X_1 + \cdots + \beta_p X_p\right) + \sigma^2.$$

De même, on peut définir un R^2 au niveau de la population par

$$R^2_{\text{pop}} = \frac{\mathrm{Var}\left(\beta_1 X_1 + \cdots + \beta_p X_p\right)}{\mathrm{Var}(Y)}$$

qui représente la fraction de variabilité de Y expliquée par le modèle, c'est-à-dire $\beta_1 X_1 + \cdots + \beta_p X_p$.

Enfin, on peut aussi ajouter l'hypothèse que l'erreur ε soit gaussienne. Dans ce cas, le modèle s'écrit

$$\left[Y \middle| X_{1:p} = x_{1:p} \right] \sim \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2 \right).$$

2 Estimation

Données : n individus $\left(x_{i,1}, x_{i,2}, \ldots, x_{i,p}, y_i \right)$ $i = 1, \ldots, n$. Conventionnellement, les données sont rangées dans un tableau où les individus sont en lignes, et les variables en colonne. On modélise les données par $\left(X_{i,1}, X_{i,2}, \ldots, X_{i,p}, Y_i \right)$. On suppose que ces n vecteurs aléatoires sont indépendants.

Ces individus étant tirés dans la population d'intérêt, on a

$$\forall i = 1, \ldots, n, \quad Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i.$$

On peut écrire matriciellement ces n équations : $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ où

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & & & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{pmatrix}.$$

2.1 Moindres carrés

L'estimateur des moindres carrés est défini par

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 = \operatorname{argmin}_{\boldsymbol{\beta}} \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right)' \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right).$$

Proposition 3. *La fonction* $\boldsymbol{\beta} \mapsto \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2$ *atteint son minimum en tout* $\widehat{\boldsymbol{\beta}}$ *qui vérifie*

$$\left(\mathbf{X}'\mathbf{X} \right) \widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}.$$

Si $\operatorname{rg}(\mathbf{X}) = p+1$, *alors* $\mathbf{X}'\mathbf{X}$ *est inversible, ce minimum est unique et on a*

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Y}.$$

Cet estimateur a les propriétés suivantes.

Théorème 4. *Sous les hypothèses qui précèdent, si* $\mathrm{rg}(\mathbf{X}) = p+1$ *p.s., on a*

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\middle|\mathbf{X}=\mathbf{x}\right) = \boldsymbol{\beta}.$$

De plus, la matrice de variance-covariance est

$$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\middle|\mathbf{X}=\mathbf{x}\right) = \sigma^2(\mathbf{x}'\mathbf{x})^{-1}.$$

En outre, conditionnellement à $\mathbf{X}=\mathbf{x}$, $\widehat{\boldsymbol{\beta}}$ *est le meilleur estimateur linéaire sans biais de* $\boldsymbol{\beta}$

Dans toute la suite de cette section, on supposera que $\mathbf{x}'\mathbf{x}$ est inversible.

S'il y a de **fortes corrélations entre covariables**, la matrice symétrique $\mathbf{x}'\mathbf{x}$ définie positive a des valeurs propres proches de 0 relativement à la somme des valeurs propres, $\mathrm{tr}(\mathbf{x}'\mathbf{x}) = n + \sum_{i=1}^{n}\sum_{j=1}^{p} x_{i,j}^2$. L'inverse $\left(\mathbf{x}'\mathbf{x}\right)^{-1}$ a donc des valeurs propres très grandes. Cela veut dire que la variabilité de l'estimateur d'un échantillon à l'autre, mesurée par la matrice de variance-covariance $\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\middle|\mathbf{X}=\mathbf{x}\right)$ devient importante. L'**estimateur est donc instable** (intervalles de confiance très larges,. . .).

On peut **comprendre ce phénomène de façon plus pragmatique** ainsi en supposant qu'il existe une combinaison linéaire des covariable presque nulle, c'est-à-dire négligeable devant σ : il existe $\lambda_0, \dots, \lambda_p$ tel que

$$\left(\lambda_0 + \lambda_1 X_1 + \cdots + \lambda_p X_p\right)^2 \ll \sigma^2$$

sur les données observées.
Alors, pour tout $a \in [-1;1]$,

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \approx (\beta_0 + a\lambda_0) + (\beta_1 + a\lambda_1) X_1 + \cdots + (\beta_p + a\lambda_p) X_p$$

dans le sens où la différence est négligeable devant σ^2. Les estimateurs $\widehat{\beta}_j$ ne savent pas s'ils doivent estimer β_j ou $\beta_j + a\lambda_j$ pour une valeur de $a \neq 0$. Ils vont **sur-apprendre** une valeur de a sur les données observées qui ne se généralise pas à l'ensemble de la population.
Nous proposerons différentes méthodes pour résoudre ce problème :

- la régression ridge : c'est la méthode qui permet de stabiliser les estimateurs lorsque les covariables sont fortement corrélées,
- la régression lasso : c'est une méthode qui permet d'annuler des coordonnées de $\widehat{\boldsymbol{\beta}}$ pour stabiliser l'estimateur,
- la sélection de variables : c'est une méthode qui permet de réduire l'ensemble des covariables.

Quant à σ^2, il est estimé avec

$$\widehat{\sigma}^2 = \frac{1}{n-(p+1)} \left\| \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \right\|_2^2.$$

Cet estimateur est sans biais dans le sens ci-dessous.

Proposition 5.

$$\mathbb{E}\left(\widehat{\sigma}^2 \middle| \mathbf{X} = \mathbf{x} \right) = \sigma^2.$$

Comme dans le cas de la régression linéaire simple, $\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ est le vecteur des **résidus**.

En cas de **sur-apprentissage**, le vecteur des résidus $\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ est de norme plus faible qu'il ne faudrait. Et l'estimateur $\widehat{\sigma}^2$ ci-dessus sous-estime la variance de l'erreur σ^2. Nous sommes alors face à une double catastrophe :

- l'estimateur $\widehat{\boldsymbol{\beta}}$ risque de pointer vers des valeurs qui ne sont **pas généralisables** à la population, et
- la variance de **l'erreur est sous-estimée**.

De plus, s'il y a de corrélation dans les colonnes de $\mathbf{x}$ au point que $\mathbf{x}'\mathbf{x}$ soit mal conditionnée, alors l'algorithme d'inversion de cette matrice est instable et,

troisième catastrophe, le calcul numérique $(\mathbf{x}'\mathbf{x})^{-1}$ n'est pas correct.

2.2 Modèle gaussien

Dans le modèle gaussien, on peut aller plus. L'équation matricielle $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ peut se ré-écrire avantageusement sous la forme

$$\left[\mathbf{Y}\middle|\mathbf{X} = \mathbf{x}\right] \sim \mathcal{N}_n\left(\mathbf{x}\boldsymbol{\beta}, \sigma^2 I_n\right),$$

où l'on n'a fait apparaître une loi gaussienne multivariée de dimension n.

La vraisemblance conditionnelle est donc donnée par :

$$L(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2\right)$$

et donc la log-vraisemblance est

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 + \text{constante.}$$

L'estimateur du maximum de vraisemblance est l'estimateur des moindres carrées pour $\boldsymbol{\beta}$, mais il est biaisé sur σ^2. On préfère donc utiliser celui introduit ci-dessus.

Théorème 6. *Sous les hypothèses gaussiennes, si* $\mathrm{rg}(\mathbf{X}) = p+1$,

$$\left[\widehat{\boldsymbol{\beta}}\middle|\mathbf{X}=\mathbf{x}\right] \sim \mathcal{N}_p\left(\boldsymbol{\beta};\ \sigma^2(\mathbf{x}'\mathbf{x})^{-1}\right).$$

De plus,

$$\left[\frac{n-(p+1)}{\sigma^2}\widehat{\sigma}^2\middle|\mathbf{X}=\mathbf{x}\right] \sim \chi^2(n-p-1).$$

Et, conditionnellement à $\mathbf{X}=\mathbf{x}$, *on a* $\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\sigma}^2$.

Ce théorème est une application directe du théorème de Cochran, dont voici une version simple.

Théorème 7 (Cochran)**.** *Soit* $\mathbf{Z} \sim \mathcal{N}_d(\boldsymbol{\mu}, \sigma^2\mathbf{I}_d)$ *un vecteur gaussien de dimension* d, *dont les coordonnées sont indépendantes et de même variance* σ^2. *Soit* $\boldsymbol{\Pi}$ *une matrice de projection orthogonale sur un sous-espace* F *de dimension* d_1 *:* $\boldsymbol{\Pi}' = \boldsymbol{\Pi}$, $\boldsymbol{\Pi}^2 = \mathbf{I}_d$, $\mathrm{tr}(\boldsymbol{\Pi}) = d_1$.
Dans ce cas, on rappelle que $(\mathbf{I}_d - \boldsymbol{\Pi})$ *est la projection orthogonale sur l'orthogonale de* F, *noté* $F^\perp$, *de dimension* $d_2 = d - d_1$.
Alors,

— *les vecteurs aléatoires* $\boldsymbol{\Pi}\mathbf{Z}$ *et* $(\mathbf{I}_d - \boldsymbol{\Pi})\mathbf{Z}$ *sont deux vecteurs gaussiens*

indépendants, de lois respectives

$$\mathbf{\Pi Z} \sim \mathscr{N}_d\Big(\mathbf{\Pi}\boldsymbol{\mu};\ \sigma^2\mathbf{\Pi}\Big), \quad (\mathbf{I}_d - \mathbf{\Pi})\mathbf{Z} \sim \mathscr{N}_d\Big((\mathbf{I}_d - \mathbf{\Pi})\boldsymbol{\mu};\ \sigma^2(\mathbf{I}_d - \mathbf{\Pi})\Big)$$

— *les variables aléatoires*

$$S_1 = \sigma^{-1}\mathbf{\Pi}\Big(\mathbf{Z} - \boldsymbol{\mu}\Big), \quad S_2 = \sigma^{-1}(\mathbf{I}_d - \mathbf{\Pi})\Big(\mathbf{Z} - \boldsymbol{\mu}\Big)$$

sont indépendantes, et de lois respectives $\chi^2(d_1)$ *et* $\chi^2(d_2)$.

Donc, la variable aléatoire

$$F = \frac{S_1 / d_1}{S_2 / d_2}$$

suit une loi de Fisher $\mathscr{F}(d_1, d_2)$.

Grâce à ce théorème, on peut obtenir des intervalles de confiance sur les coordonnées de $\widehat{\boldsymbol{\beta}}$. Comme celles-ci sont corrélées, il est aussi intéressant de considérer l'ellipsoïde de confiance. On pose

$$\mathbf{v} = \big(\mathbf{x}'\mathbf{x}\big)^{-1}$$

de telle sorte que, la variance de $\sigma^{-1}\widehat{\beta}_j$ est donnée par le j-ième coefficient de la diagonale de cette matrice v_{jj}.

Proposition 8. *Conditionnellement à* $\mathbf{X}=\mathbf{x}$*, l'intervalle centré en* $\widehat{\beta}_j$*, de bornes*

$$\widehat{\beta}_j \pm u\,\widehat{\sigma}\sqrt{v_{jj}}$$

où u *est le quantile d'ordre* $1-\alpha/2$ *de la loi de Student* t_{n-p-1} *est un intervalle de confiance de niveau* $(1-\alpha)$ *pour* β_j*. Autrement dit,*

$$\mathbb{P}\left(\beta_j \in \left[\widehat{\beta}_j \pm u\,\widehat{\sigma}\sqrt{v_{jj}}\right] \middle| \mathbf{X}=\mathbf{x}\right) = 1-\alpha.$$

Conditionnellement à $\mathbf{X}=\mathbf{x}$*, l'ellipsoïde « centré en »* $\widehat{\boldsymbol{\beta}}$*, défini par*

$$\mathscr{E} = \left\{\mathbf{b}\in\mathbb{R}^{p+1} \ : \ \left\|\mathbf{x}\left(\mathbf{b}-\widehat{\boldsymbol{\beta}}\right)\right\|_2^2 \leq u\,(p+1)\widehat{\sigma}^2\right\}$$

où u *est le quantile d'ordre* $1-\alpha$ *de la loi de Fisher* $\mathscr{F}(p+1, n-p-1)$ *est une région de confiance de niveau* $(1-\alpha)$ *pour* $\boldsymbol{\beta}$*. Autrement dit,*

$$\mathbb{P}\left(\boldsymbol{\beta}\in\mathscr{E} \middle| \mathbf{X}=\mathbf{x}\right) = 1-\alpha.$$

Il faut noter ici que :

— le « centre » de l'ellipsoïde $\widehat{\boldsymbol{\beta}}$ est aléatoire,

— la forme de l'ellipsoïde est donnée par la matrice $\mathbf{x}$, donc les corrélations entre les covariables,
— le « rayon au carré » $u(p+1)\widehat{\sigma}^2$ est lié à l'estimation de l'erreur, et est donc aléatoire.

De façon plus générale, on peut s'intéresser à la région de confiance du vecteur $\boldsymbol{\theta} = \mathbf{T}\boldsymbol{\beta}$, où $\mathbf{T}$ est une matrice $q \times (p+1)$, $q \leq (p+1)$, de rang q. Ici, $\boldsymbol{\theta}$ est donc un vecteur de dimension q.

Proposition 9. *Sous les hypothèses qui précèdent, l'ellipsoïde « centré en »* $\mathbf{T}\widehat{\boldsymbol{\beta}}$*, défini par*

$$\mathscr{E}_{\mathbf{T}} = \left\{ \mathbf{b} \in \mathbb{R}^{p+1} \ : \ (\mathbf{T}\mathbf{b} - \mathbf{T}\widehat{\boldsymbol{\beta}})' \left(\mathbf{T}(\mathbf{x}\mathbf{x}')^{-1}\mathbf{T}' \right)^{-1} (\mathbf{T}\mathbf{b} - \mathbf{T}\widehat{\boldsymbol{\beta}}) \leq u\, q \widehat{\sigma}^2 \right\}$$

où u *est le quantile d'ordre* $1-\alpha/2$ *de la loi de Ficher* $\mathscr{F}(q, n-p-1)$ *est une région de confiance au niveau* $(1-\alpha)$ *pour* $\mathbf{T}\boldsymbol{\beta}$*.*

De ces intervalles et régions de confiance, on peut en déduire des tests statistiques (qui sont très souvent inutiles). Par exemple, pour $j \in \{0, 1, \ldots, p\}$ fixé, la décision au niveau $(1-\alpha)$ du test :

$$H_0 : \beta_j = 0, \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

s'obtient en choisissant :

— de conserver H_0 si 0 est dans l'intervalle de confiance de niveau $1-\alpha$,

— de rejeter H_0 sinon.

Ce sont les t-tests de nullité des coefficients.

On peut également réalisation un test pour décider entre

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{vs} \quad H_1 : \text{il existe } j \text{ entre } 1 \text{ et } p \text{ tel que } \beta_j \neq 0.$$

C'est le F-test d'intérêt du modèle. Pour cela, on doit utiliser les résultats de la proposition 9 avec $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p)$, $q = p$. Dans ce cas, la matrice $\mathbf{T}$ est donnée par

$$\mathbf{T} = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

ce qui permet de se débarrasser de β_0. Dans le test au niveau $(1-\alpha)$, on décide :

— de conserver H_0 si $0 \in \mathscr{E}_{\mathbf{T}}$, i.e. si 0 est dans l'ellipsoïde de confiance de niveau $(1-\alpha)$,

— de rejeter H_0 sinon.

3 Prédictions

Le modèle linéaire a deux objectifs.

1. **Comprendre et expliquer la variabilité de Y grâce aux covariables.** Dans ce cas, on ajuste un modèle linéaire et on interprète
   — les effets (β_j, $j \neq 0$),
   — la variabilité quand on fixe toutes les covariables (σ^2) et
   — le pourcentage de variabilité de Y expliqué par le modèle (R^2)...
2. **Prédire la valeur de Y pour un nouvel individu.** Dans ce cas, on se donne de nouvelles valeurs des covariables pour un individu : $\widetilde{x}_1, \ldots, \widetilde{x}_p$. Et on veut :
   — estimer la moyenne de Y sur la sous-population où les covariables sont fixées à $\widetilde{\mathbf{x}}$, à savoir le paramètre $\theta = \beta_0 + \beta_1 \widetilde{x}_1 + \cdots + \beta_p \widetilde{x}_p$,
   — prédire une valeur de la réponse, notée $\widetilde{Y}$, pour un individu fixé, issu de cette sous-population.

Ces deux objectifs peuvent être mener simultanément. Souvent, l'un des deux objectifs est privilégié. Le premier objectif est lié à des problèmes d'estimation, d'erreur d'inférence, entièrement traité dans ce qui précède via les variances des

estimateurs ou les intervalles/régions de confiance. On se concentre maintenant sur le second objectif. On notera que le paramètre θ, de dimension 1, est de la forme $\mathbf{T}\boldsymbol{\beta}$, où $\mathbf{T}$ est la matrice ligne

$$\mathbf{T} = \begin{pmatrix} 1 & \widetilde{x}_1 & \widetilde{x}_2 & \cdots & \widetilde{x}_p \end{pmatrix}.$$

On se place maintenant sous les hypothèses du **modèle linéaire gaussien** de la section 2. La proposition 9 permet de donner un intervalle de confiance pour θ, en utilisant la matrice ligne $\mathbf{T}$ introduite ci-dessus.

Dans ce cas, on utilise l'estimateur $\widehat{\theta} = \mathbf{T}\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1\widetilde{x}_1 + \cdots + \widehat{\beta}_p\widetilde{x}_p$. Et on peut montrer que l'intervalle de confiance pour θ au niveau $(1-\alpha)$ est un intervalle centré en $\widehat{\theta}$, dont les bornes sont

$$\widehat{\theta} \pm u\widehat{\sigma}\sqrt{\mathbf{T}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{T}'},$$

où u est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-p-1$ degrés de liberté.

Pour la prédiction de $\widetilde{Y}$, on obtient un intervalle de prédiction de niveau $(1-\alpha)$

avec l'intervalle centré en $\widehat{\theta}$, dont les bornes sont

$$\widehat{\theta} \pm u\widehat{\sigma}\sqrt{1+\mathbf{T}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{T}'},$$

où u est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-p-1$ degrés de liberté. Le terme $1+$ qui apparaît en rouge sous la racine carré vient du fait que l'on doit tenir compte de la variabilité individuelle d'un individu à l'autre dans la sous-population où les covariables sont fixées à $\widetilde{\mathbf{x}}$. Cette variabilité individuelle est donnée par σ^2 (en variance), estimée par $\widehat{\sigma}^2$, et s'ajoute à la variabilité de l'estimateur, car le nouvel individu n'est pas dans l'échantillon des données.

4 Des covariables transformées

Variables catégorielles On ne peut pas ajouter directement dans une formule du type $\beta_0+\beta_1 X_1+\cdots+\beta_p X_p$ une variable catégorielle. Lorsque l'on veut utiliser une variable catégorielle, il faut la transformer en une ou plusieurs **variables binaires** (=Dont les seules valeurs possibles sont 0 et 1).

Lorsqu'il s'agit d'une variable catégorielle à **deux modalités** (homme - femme ;

diplômée - non diplômée ;...), on définit une variable X_j qui vaut 1 pour une modalité (ex : femme), et 0 pour l'autre modalité (ex : homme). Dans la formule linéaire, le terme $+\beta_j X_j$ peut donc prendre deux valeurs 0 ou β_j. On interprète alors β_j comme l'écart de moyenne entre les sous-populations pour la modalité codée par 1 (toutes les autres covariables étant fixées), et la modalité codée par 0 (toutes les autres covariables étant fixées aux mêmes valeurs). La modalité codée par 0 est appelée **modalité de référence**.

Lorsque la variable a K modalités, il faut introduire $K-1$ variables binaires :

- on fixe une modalité de référence a_0 parmi les K modalités,
- pour chacune des modalités a différentes de la modalité de référence, on introduit une variable binaire égale à 1 si cette modalité a est celle observée, 0 sinon.

Ces $(K-1)$ variables binaires portent la même information que la variable catégorielle à K modalités et,

- soit elles sont toutes nulles (c'est la modalité de référence qui est celle observée)
- soit une et une seule d'entre elle est égale à 1, les autres sont égales à 0.

On peut alors interpréter les coefficients en facteur de ces covariables binaires comme des différences de moyennes entre deux sous-populations, où, toutes les

autres covariables différentes de ces $K-1$ variables binaires sont fixées, et l'une des deux sous-population correspond à la modalité de référence a_0, l'autre à la modalité liée à la covariable binaire a.

Interaction L'une des propriétés importantes des modèles linéaires est l'interprétation de β_j comme effet lorsqu'on augmente X_j d'une unité, sachant toutes les autres covariables fixées. Cet effet ne dépend pas des valeurs des valeurs des autres covariables. On peut chercher à rendre l'effet dépendant des autres covariables. Pour rester linéaire, si on veut faire dépendre cet effet de la valeur de X_k, on va

$$\text{remplacer} \quad \beta_j \quad \text{par} \quad \beta_j + \beta_{j,k} X_k.$$

Dans la formule linéaire, en reportant, on obtient

$$\begin{aligned}\beta_0 + \beta_1 X_1 + \cdots + (\beta_j + \beta_{j,k} X_k) X_j + \cdots + \beta_p X_p \\ = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{j,k} X_k X_j.\end{aligned}$$

Cette opération revient donc à ajouter une nouvelle covariable, qui est le produit de deux covariables initiales, dans la formule linéaire. De telles nouvelles covariables s'appellent des interactions, et le nouveau coefficient en facteur $\beta_{j,k}$

s'interprète comme expliqué au dessus. (Notons que X_j et X_k jouent des rôles symétriques).
Attention, il ne faut jamais introduire une variable d'interaction (c'est-à-dire une variable produit) sans avoir utiliser les deux covariables dans le modèle. Sinon, on peut prendre pour une interaction ce qui serait en fait un effet direct de la covariable manquante. . .

Autres transformations non linéaires On peut introduire de nouvelles covariables à partir des anciennes en appliquant des transformations non linéaires aux covariables initiales. Par exemple, s'il n'y avait qu'une seule covariables X_1, on peut introduire les nouvelles covariables

$$X_2 = (X_1)^2,\ X_3 = (X_1)^3, \ldots, X_p = (X_1)^p.$$

Dans ce cas, la formule linéaire

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \beta_0 + \beta_1 X_1 + \beta_2 (X_1)^2 + \cdots + \beta_p (X_1)^p$$

est **un polynôme de degré p en X_1**. Mais les coefficients n'ont plus une interprétation simple. . . Le problème des polynômes de degré p est leur explosion vers $+\infty$ ou $-\infty$ à la vitesse $(X_1)^p$ quand $|X_1| \to +\infty$.

Plutôt qu'un polynôme, on peut utliiser une fonction **affine par morceau qui soit continue**, et où les jointures ont lieu en $\xi_1, \xi_2, \ldots, \xi_K$. Cela revient à introduire les nouvelles covariables :

$$X_2 = (X_1 - \xi_1)_+, \ldots, X_{K+1} = (X_1 - \xi_K)_+,$$

où, pour tout $u \in \mathbb{R}$, u_+ désigne la partie positive de u, c'est-à-dire $u_+ = \max(0, u)$. Notons que les ξ_k sont les points de discontinuité de la dérivée d'ordre 1.

Plutôt qu'une fonction affine par morceaux, on peut utiliser un polynôme par morceau de degré 3, C^2 sur tout $\mathbb{R}$, avec des discontinuité dans la dérivée d'ordre 3. De ce cas, cela revient à introduire dans le modèle linéaireles covariables :

$$X_2 = (X_1)^2,\ X_3 = (X_1)^3, X_4 = (X_1 - \xi_1)^3_+, \ldots, X_{K+3} = (X_1 - \xi_K)^3_+$$

On parle alors de **spline cubique**... Notons que les coefficients en facteur de ces covariables ne s'interprètent pas non plus.

Les splines cubiques dites naturelles sont des variantes de la situation ci-dessous où, sur les deux intervalles infinis $]-\infty; \xi_1[$ et $]\xi_K; +\infty[$, on impose que la fonction soit linéaire pour ne pas exploser trop vite vers l'infini. Cela ajoute $2 \times 2 = 4$ contraintes d'égalité sur les coefficients $\beta_1, \ldots, \beta_{K+3}$.